Fitting Finite Mixtures of Linear Regression Models with Varying & Fixed Effects in R*

Bettina Grün[1] and Friedrich Leisch[2]

[1] Department of Statistics and Probability Theory
Vienna University of Technology
Wiedner Hauptstraße 8-10/1071
A-1040 Wien, Austria
Bettina.Gruen@ci.tuwien.ac.at

[2] Department of Statistics
University of Munich
Ludwigstraße 33
D-80539 München, Germany
Friedrich.Leisch@stat.uni-muenchen.de

Summary. A general model class of finite mixtures of linear regression models is presented. It allows (nested) varying and fixed effects for the regression coefficients and the variance. A combination of varying and fixed effects is useful in applications because it can be used to account for overdispersion as a nuisance parameter or to reduce the number of estimated parameters. In addition concomitant variable models for the component weights provide the possibility to partition the data into the mixture components through other variables.

Maximum likelihood parameter estimation using the EM algorithm is outlined and the implementation in R by extending package **flexmix** is described. In this paper multinomial logit concomitant variable models are considered, but the provided infrastructure allows to easily define new concomitant models and rapid prototyping is possible if functionality already available in R can be used.

Key words: concomitant variables, finite mixture models, linear regression, R, unobserved heterogeneity

1 Introduction

Finite mixtures of regression models are frequently used in applications to model unobserved heterogeneity. Areas of application include biology, economics, marketing

* This paper was published as: Bettina Grün and Friedrich Leisch. Fitting finite mixtures of linear regression models with varying & fixed effects in R. In Alfredo Rizzi and Maurizio Vichi, editors, Compstat 2006—Proceedings in Computational Statistics, pages 853–860. Physica Verlag, Heidelberg, Germany, 2006.

and medicine [WCP98; Ait99; MP00]. The unobserved heterogeneity in the data might either be considered a nuisance parameter which has to be accounted for as, e.g. in random intercept models [FL89], or the main interest is to detect underlying groups in the data, as e.g. in market segmentation [WK01]. In both cases it might be desirable to restrict certain coefficients to be equal over all components as this reduces the number of estimated parameters, or the equality is a-priori assumed due to expert knowledge. In the following we refer to the fixed parameters as fixed effects and to the others, which vary between the components, as varying effects. In addition we also introduce varying effects between groups of components where equality of the coefficients is assumed for each group.

In model-based clustering, where the dependent variable has a multivariate Gaussian distribution and the independent variables consist only of the intercept, different models to restrict the dispersion parameters over the components have been proposed which take the orientation, shape and volume of the variance-covariance matrices into account [FR02]. It is therefore natural to also consider restrictions for the variance parameters of the components for mixtures of linear regression models.

In order to characterize the different components of the mixture the use of concomitant variables has been suggested [DM88]. Different concomitant variable models are possible, but multinomial logit models are the most popular ones [WCP98].

In Section 2 the model class of finite mixtures of linear regression models with varying and fixed effects is introduced and the maximum likelihood estimation of the parameters using the EM algorithm [DLR77] is outlined. Section 3 presents the model fitting in R [R D05] using package **flexmix** [Lei04]. The implementational details which are especially of interest if the functionality shall be extended, e.g. by specifying a different concomitant variable model, are given in Section 4.

2 Model specification

Finite mixtures of regression models are given by

$$H(y|\boldsymbol{x}, \boldsymbol{w}, \Theta) = \sum_{k=1}^{K} \pi_k(\boldsymbol{w}, \boldsymbol{\alpha}) F_k(y|\boldsymbol{x}, \boldsymbol{\vartheta}_k)$$

where Θ denotes the vector of all parameters. The dependent variables are y, the independent $\boldsymbol{x}$ and the concomitant $\boldsymbol{w}$. F_k is the component specific distribution function. The component specific parameters are given by $\boldsymbol{\vartheta}_k$. For the component weights π_k it holds $\forall \boldsymbol{w}$ that

$$\sum_{k=1}^{K} \pi_k(\boldsymbol{w}, \boldsymbol{\alpha}) = 1 \quad \wedge \quad \pi_k(\boldsymbol{w}, \boldsymbol{\alpha}) \geq 0 \, \forall k \tag{1}$$

where $\boldsymbol{\alpha}$ are the parameters of the concomitant variable model.

In the following only finite mixtures where the component specific distribution functions follow the Gaussian distribution are considered, i.e. $F_k \equiv \Phi$. In this case the component specific parameters are given by $\boldsymbol{\vartheta}_k = (\boldsymbol{\beta}_k', \sigma_k^2)$ where $\boldsymbol{\beta}_k$ are the regression coefficients and σ_k^2 denotes the variance.

Different concomitant variable models are possible to determine the component weights [DM88]. The mapping function only has to fulfill condition (1). In the following we assume a multinomial logit model for the π_k given by

$$\pi_k(\boldsymbol{w}, \boldsymbol{\alpha}) = \frac{e^{\boldsymbol{w}'\boldsymbol{\alpha}_k}}{\sum_{u=1}^{K} e^{\boldsymbol{w}'\boldsymbol{\alpha}_u}} \quad \forall k$$

with $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_k')_{k=1,\ldots,K}$ and $\boldsymbol{\alpha}_1 \equiv \mathbf{0}$.

The component specific parameters $\boldsymbol{\vartheta}_k$ are either restricted to be equal over all components, to vary between groups of components or for all components. The varying between groups is referred to as varying effects with one level of nesting. For the nesting a disjoint partition K_c, $c \in C$ of the set $\tilde{K} := \{1 \ldots, K\}$ is defined for the regression coefficients which are accordingly split into three groups:

$$\boldsymbol{\beta}_k = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_{2,c(k)}', \boldsymbol{\beta}_{3,k}')'$$

where $c(k) = \{c \in C : k \in K_c\}$. Similar a disjoint partition K_v, $v \in V$, of $\tilde{K}$ is defined for the variance parameters which gives:

$$\sigma_k^2 = (\sigma_1^2, \sigma_{2,v(k)}^2, \sigma_{3,k}^2)$$

where $v(k) = \{v \in V : k \in K_v\}$.

2.1 Parameter estimation

For maximum likelihood estimation of finite mixture models with a fixed number of components K, the EM algorithm is the most popular solution. The unobserved component memberships z of the observations are treated as missing values and the data are augmented by estimates of the component memberships, i.e. the estimated a-posteriori probabilities $\hat{p}_{nk}$. For a sample of N observations $\{(y_1, \boldsymbol{x}_1, \boldsymbol{w}_1), \ldots, (y_N, \boldsymbol{x}_N, \boldsymbol{w}_N)\}$ the EM-algorithm is given by:

E-step: Given the current parameter estimates $\boldsymbol{\Theta}^{(i)}$ in the i-th iteration, replace the missing data z by the estimated a-posteriori probabilities

$$\hat{p}_{nk} = P(k|y_n, \boldsymbol{x}_n, \boldsymbol{w}_n, \boldsymbol{\Theta}^{(i)}) = \frac{\pi_k(\boldsymbol{w}_n, \boldsymbol{\alpha}^{(i)}) F(y_n|\boldsymbol{x}_n, \boldsymbol{\vartheta}_k^{(i)})}{\sum_{u=1}^{K} \pi_u(\boldsymbol{w}_n, \boldsymbol{\alpha}^{(i)}) F(y_n|\boldsymbol{x}_n, \boldsymbol{\vartheta}_u^{(i)})}$$

M-step: Given the estimates for the a-posteriori probabilities $\hat{p}_{nk}$ (which are functions of $\boldsymbol{\Theta}^{(i)}$), obtain new estimates $\boldsymbol{\Theta}^{(i+1)}$ of the parameters by maximizing

$$Q(\boldsymbol{\Theta}^{(i+1)}|\boldsymbol{\Theta}^{(i)}) = Q_1(\boldsymbol{\vartheta}^{(i+1)}|\boldsymbol{\Theta}^{(i)}) + Q_2(\boldsymbol{\alpha}^{(i+1)}|\boldsymbol{\Theta}^{(i)})$$

where

$$Q_1(\boldsymbol{\vartheta}^{(i+1)}|\boldsymbol{\Theta}^{(i)}) = \sum_{n=1}^{N} \sum_{k=1}^{K} \hat{p}_{nk} \log(F(y_n|\boldsymbol{x}_n, \boldsymbol{\vartheta}_k^{(i+1)})) \tag{2}$$

and

$$Q_2(\boldsymbol{\alpha}^{(i+1)}|\boldsymbol{\Theta}^{(i)}) = \sum_{n=1}^{N} \sum_{k=1}^{K} \hat{p}_{nk} \log(\pi_k(\boldsymbol{w}_n, \boldsymbol{\alpha}^{(i+1)})). \tag{3}$$

Q_1 and Q_2 can be maximized separately. The maximization of Q_1 gives new estimates $\boldsymbol{\vartheta}^{(i+1)}$ and the maximization of Q_2 gives $\boldsymbol{\alpha}^{(i+1)}$. Q_1 is maximized using weighted ML estimation of linear models and Q_2 using weighted ML estimation of multinomial logit models.

If there are only varying effects for the component-specific parameters specified they can be determined separately for each component. If there are also fixed or nested varying effects, the vector of observations $\boldsymbol{y} = (y_n)_{n=1,\ldots,N}$ has to be replicated K times and the covariate matrix $\boldsymbol{X} = (\boldsymbol{X}_{\text{fixed}}, \boldsymbol{X}_{\text{nested}}, \boldsymbol{X}_{\text{varying}})$ is given by

$$\begin{aligned}
\boldsymbol{X}_{\text{fixed}} &= \boldsymbol{1}_K \otimes (\boldsymbol{x}'_{1,n})_{n=1,\ldots,N} \\
\boldsymbol{X}_{\text{nested}} &= \boldsymbol{J} \odot (\boldsymbol{x}'_{2,n})_{n=1,\ldots,N} \\
\boldsymbol{X}_{\text{varying}} &= \boldsymbol{I}_K \otimes (\boldsymbol{x}'_{3,n})_{n=1,\ldots,N}
\end{aligned}$$

$\boldsymbol{1}_K$ is a vector of 1s of length K. $\boldsymbol{J}$ is the incidence matrix for each component $k = 1, \ldots, K$ and each nesting group $c \in C$ and hence is of dimension $K \times |C|$. $\boldsymbol{I}_K$ is the identity matrix of dimension $K \times K$. $\otimes$ denotes the Kronecker product, $\odot$ represents the Khatri-Rao product (i.e. the column-wise Kronecker product) and $\boldsymbol{x}_{m,n}$ are the covariates which correspond to the coefficients $\boldsymbol{\beta}_{m,\cdot}$ for $m = 1, 2, 3$.

The EM algorithm is initialized by randomly assigning a-posteriori probabilities to each observation, e.g. by partitioning the data into the components using a cluster algorithm. After each M-step the convergence of the algorithm is checked. The algorithm is stopped if the change in the likelihood is smaller than a pre-specified ϵ or the maximum number of iterations is reached. Furthermore, after each M-step the average component sizes (over the given data points) are checked and components which are smaller than a given (relative) minimum size are omitted in order to avoid too small components where fitting problems might arise.

It has been shown that the values of the likelihood are monotonically increased during the EM algorithm. On the one hand this ensures the convergence of the EM algorithm if the likelihood is bounded, but on the other hand only the detection of a local maximum can be guaranteed. Therefore, it is in general recommended to repeat the EM algorithm with different initializations and choose as final solution the one with the maximum likelihood.

3 Using the new functionality

We illustrate the fitting of a finite mixture of the model class presented in Section 2 on a simple artificial example. For each component we assume a linear regression model with two covariates and an intercept. There are varying effects for the intercept, nested varying effects for variable x_1 and fixed effects for x_2. The mixture has three components and the regression model for each component is given by

$$\begin{aligned}
&\text{Class 1: } y = -8 + 10x_1 + 5x_2 + \epsilon \\
&\text{Class 2: } y = 1 + 10x_1 + 5x_2 + \epsilon \\
&\text{Class 3: } y = 3 + 5x_2 + \epsilon
\end{aligned}$$

with $\epsilon \sim N(0, 1)$.

The component weights depend on the variable w and are determined by

$$\begin{aligned}
&\text{Class 2: } \text{logit}(\pi_2(w, \boldsymbol{\alpha})) = 2 - 2w \\
&\text{Class 3: } \text{logit}(\pi_3(w, \boldsymbol{\alpha})) = 2w
\end{aligned}$$

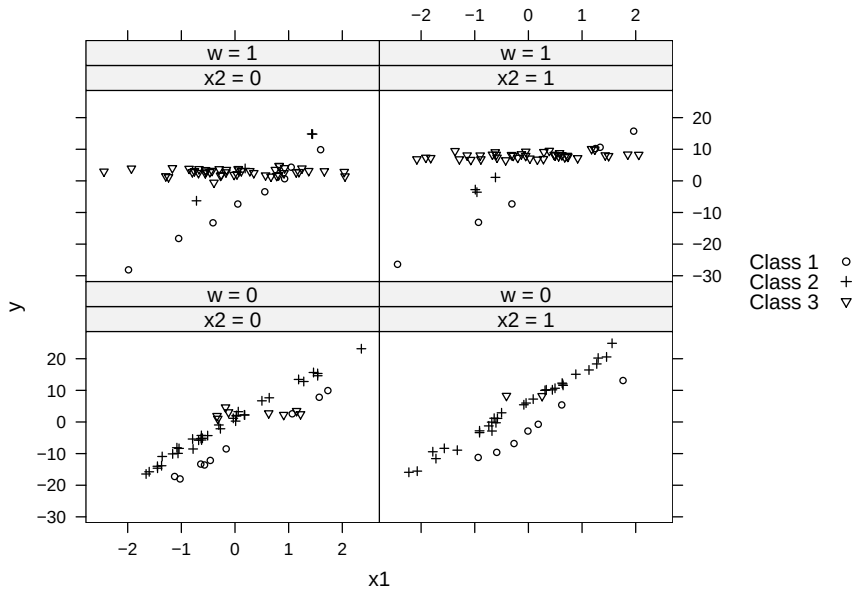

Fig. 1. Sample with 200 observations from the artificial example.

The covariates x_1 and x_2 and the concomitant variable w are mutually independent. x_1 is sampled from the standard Gaussian distribution. x_2 and w are each sampled from $\{0, 1\}$ with equal probability. In Figure 1 a sample with 200 observations from this mixture is shown. The effect of the concomitant variable can clearly be seen as for $w = 0$ class 2 is the largest with $\pi = 0.79$ and for $w = 1$ class 3 also with $\pi = 0.79$.

The model has been implemented in R package **flexmix** (available from `http://cran.R-project.org`), see [Lei04] for details on using the software. After loading the package itself and the data set `NregFix`, the component specific model is specified using function `FLXglmFix`:

```
> Model <- FLXglmFix(fixed = ~x2, nested = list(k = c(2,
+     1), formula = c(~x1, ~0)), varFix = TRUE)
```

Compared to `FLXglm`, which is already available in **flexmix** to specify component models where only varying effects are allowed, there are additional arguments for the fixed and nested varying effects of the regression coefficients and the variance.

We fit the model 5 times with the EM algorithm and return the best model with respect to the log-likelihood. For the concomitant variable model we use function `FLXmultinom`:

```
> fittedModel <- stepFlexmix(y ~ 1, model = Model,
+     nrep = 5, data = NregFix, concomitant = FLXmultinom(~w))

> fittedModel
```

```
Call:
stepFlexmix(y ~ 1, model = Model, data = NregFix,
    concomitant = FLXmultinom(~w), nrep = 5)

Cluster sizes:
 1  2  3
32 74 94

convergence after 20 iterations
```

The returned object is a `flexmix` object and therefore, the already available functions for inspecting the fitted model such as e.g. `plot`, `summary` or `parameters` can be used.

The fitted coefficients plus tests for significance are obtained with

```
> SummaryFittedModel <- summary(refit(fittedModel))
```

The coefficients are ordered such that the fixed coefficients are first, the nested varying coefficients second and the varying coefficients last.

4 Implementational details

The implementation extends the R package **flexmix** and tries to conform to the main aims of the package, which is easy extendibility and fast prototyping for new types of mixture models. It uses S4 classes and methods [Cha98] as implemented in the R package **methods**.

A new M-step driver is provided which fits finite mixtures of linear regression models with fixed and nested varying effects for the coefficients and the variance. The class `"FLXglmFixmodel"` returned by the driver `FLXglmFix` has the additional slots with respect to `"FLXglmmodel"`:

design: An incidence matrix indicating which columns of the model matrix are used for which component.

nestedformula: An object of class `"FLXnested"` containing the formula for the nested effects of the regression coefficients and the number of components in each K_c, $c \in C$.

fixed: The formula for the fixed effects of the regression coefficients.

variance: A logical indicating if varying effects should be estimated or a vector specifying the grouping of the nested effects for the variance.

The difference between estimating finite mixtures including only varying effects using models specified with `FLXglm` and those with varying and fixed effects using function `FLXglmFix` is hidden from the user, as the user interface for the function `flexmix` is the same. The fitted model is of class `"flexmix"` and can be analyzed using the same functions as for any model fitted using package **flexmix**. The methods used are the same except if the slot containing the model is accessed and method dispatching is made via the model class. New methods are provided for models of class `"FLXglmFixmodel"` for the following functions: `flexmix`, the fitter function `FLXfit` and the functions `refit`, `fitted` and `predict` which can be used for analyzing the fitted model.

The implementation allows repeated measurements by specifying a grouping variable in the formula argument of the `flexmix` call. However, multiple independent responses are not possible at the moment. Furthermore, it has to be noticed that the formulas of the different effects are evaluated by updating the formula of the random effects successively with the formula of the fixed and then of the nested varying effects. This ensures that if a random effect is fitted to the intercept, the model matrix of a factor includes only the remaining columns for the fixed effects to have full column rank. However, this updating scheme makes it impossible to estimate fixed effects for the intercept while fitting random effects to a factor.

For representing concomitant variable models the class `"FLXconcomitant"` is defined. It specifies how the concomitant model is fitted and has the following slots:

fit: A function (x, y, ...) returning the fitted values for the component weights during the EM algorithm.
df: A function (x, k, ...) returning the degrees of freedom used for estimating the concomitant model given the model matrix `x` and the number of components `k`.
x: A matrix containing the data of the concomitant variables.
formula: Formula for determining the model matrix `x`.
name: A character string describing the model.

Two constructor functions for concomitant variable models are provided. `FLXconstant` is for constant component weights without concomitant variables and `FLXmultinom` for multinomial logit models. `FLXmultinom` has its own class which extends `"FLXconcomitant"` and has additional slots for the fitted coefficients and a refit function. The multinomial logit models in equation (3) are fitted using package **nnet** [VR02].

5 Summary & outlook

A general model class for fitting finite mixtures of linear regression models has been presented. The regression coefficients can be modelled using fixed and (nested) varying effects. A similar structure is possible for the variances. For characterizing the components concomitant variables are introduced where the component membership is determined using multinomial logit models. The parameter estimation is outlined and the implementation in R extending package **flexmix** is presented and its application demonstrated on an artificial example.

In the future we want to extend this model class to allow for generalized linear models [MN89] in the components. For this purpose, other distributions have to be possible for the dependent variable. Finally, the same functionality as `FLXglm` shall be provided for varying and fixed effects. Furthermore, it would be interesting to consider other concomitant variable models, as e.g. multinomial probit models or constant component memberships for groups induced by the concomitant variables. This extension should be easily possible by providing a new constructor function for the concomitant variable model represented by an object of class `"FLXconcomitant"` which specifies the fitting function given the concomitant variable model matrix and the a-posteriori probabilities.

In addition, it would be interesting to consider different estimation methods. For example the use of vertex direction algorithms [Böh95] to provide good initial values for the EM algorithm could be investigated.

Acknowledgement. This research was supported by the Austrian Academy of Sciences (ÖAW) through a DOC-FFORTE scholarship for Bettina Grün and the Austrian Science Foundation (FWF) under grant P17382.

References

[Ait99] M. Aitkin. A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, 55:117–128, 1999.

[Böh95] D. Böhning. A review of reliable maximum likelihood algorithms for semiparametric mixture models. *Journal of Statistical Planning and Inference*, 47:5–28, 1995.

[Cha98] J. M. Chambers. *Programming with Data*. Springer, New York, 1998.

[DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM-algorithm. *Journal of the Royal Statistical Society B*, 39:1–38, 1977.

[DM88] C. M. Dayton and G. B. Macready. Concomitant-variable latent-class models. *Journal of the American Statistical Association*, 83(401):173–178, March 1988.

[FL89] D. A. Follmann and D. Lambert. Generalizing logistic regression by nonparametric mixing. *Journal of the American Statistical Association*, 84:295–300, 1989.

[FR02] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, June 2002.

[Lei04] F. Leisch. FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, 11(8), 2004.

[MN89] P. McCullagh and J. A. Nelder. *Generalized linear models*. Chapman and Hall, 1989.

[MP00] G. J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, 2000.

[R D05] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005. ISBN 3-900051-07-0.

[VR02] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002.

[WCP98] P. Wang, I. M. Cockburn, and M. L. Puterman. Analysis of patent data — A mixed-poisson-regression-model approach. *Journal of Business & Economic Statistics*, 16(1):27–41, 1998.

[WK01] M. Wedel and W. A. Kamakura. *Market Segmentation — Conceptual and Methodological Foundations*. Kluwer Academic Publishers, 2nd edition, 2001.